COMPYSYS 726

Dataset Description

Note:

There are two zip files : datasets.zip and mnist\_fmnist.zip. The datasets.zip is a collection of tiny datasets and should be useful on most computers. The dataset.zip file includes a few more that the list detailed below. I suggest that you use the five listed in this document.

The mnist and fashion mnist data are much bigger and you will need to use a more capable computer. These are better suited for exposing the issues of overfitting

The original datasets have been obtained from various repositories: UCI, KAGGEL and others. These have not been curated. The original data is in the CSV files. After importing the CSV files, you will need to -

* Rearrange the matrices into the correct shape to suit Matlab or Python : [nFeatures x nPats] or [nPats x nFeatures]
* Scale and normalize the means and StdDev of each feature
* One hot encode the output classes
* Split the data into three sets: training (70%), validation (15%), test (15%), or training (70%), validation (30%),

A Matlab ‘mat’ file may also be provided with the outputs one-hot-encoded.

1. Iris Dataset:

irisInputs - a 4x150 matrix of four attributes of 150 flowers.

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

irisTargets - a 3x150 matrix of 150 associated class vectors defining which of three classes each input is assigned to. Classes are represented by a 1 in one of three rows, with zeros in the others.

1. Cancer:

cancerInputs - a 9x699 matrix defining nine attributes of 699 biopsies.

1. Clump thickness
2. Uniformity of cell size
3. Uniformity of cell shape
4. Marginal Adhesion
5. Single epithelial cell size
6. Bare nuclei
7. Bland chomatin
8. Normal nucleoli
9. Mitoses

cancerTargets - a 2x699 matrix where each column indicates a correct category with a one in either element 1 or element 2.

1. Benign
2. Malignant
3. Glass:

glassInputs - a 9x214 matrix of nine attributes of 214 glass samples

1. Refractive index
2. Sodium (unit measurement: weight percent in corresponding oxide)
3. Magnesium
4. Aluminum
5. Silicon
6. Potassium
7. Calcium
8. Barium
9. Iron

glassTargets - a 2x214 matrix where each column indicates a correct category with a one in either element 1 or element 2.

1. Window glass
2. Non-window glass

4. Thyroid:

thyroidInputs - a 21x7200 matrix consisting of 7200 patients characterized by 15 binary and 6 continuous patient attributes. thyroidTargets - a 3x7200 matrix of 7200 associated class vectors defining which of three classes each input is assigned to. Classes are represented by a 1 in row 1, 2 or 3.

1. Normal, not hyperthyroid
2. Hyperfunction
3. Subnormal functioning

5. Wine:

wineInputs - a 13x178 matrix of thirteen attributes of 178 wines.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

wineTargets - a 3x178 matrix of 7200 associated class vectors defining which of three classes each input is assigned to. Classes are represented by a 1 in row 1, 2 or 3.

1. Vinyard #1
2. Vinyard #2
3. Vinyard #3